# Top-label calibration

**Chirag Gupta** [1] **Aaditya Ramdas** [1]

## Abstract

We study the problem of post-hoc calibration for multiclass classification, with an emphasis on histogram binning. Multiple works have focused on calibration with respect to the confidence of just the predicted class (or 'top-label'). We find that the popular notion of confidence calibration (Guo et al., 2017) is not sufficiently strong—there exist predictors that are not calibrated in any meaningful way but are perfectly confidence calibrated. We propose a closely related (but subtly different) notion, *top-label calibration*, that accurately captures the intuition and simplicity of confidence calibration, but addresses its drawbacks. We formalize a histogram binning (HB) algorithm that reduces top-label multiclass calibration to the binary case and explore its practical performance in the post-hoc calibration setting. We also formalize an HB algorithm corresponding to the stricter notion of class-wise calibration. In experiments with deep neural nets, we find that our principled versions of HB are often better than temperature scaling, for both top-label and class-wise calibration. Code for this work will be open sourced at `https://github.com/aigen/df-posthoc-calibration`.

## 1. Introduction

Calibration is a desirable property of validity for probabilistic predictions on a categorical outcome. For example, consider a meteorologist who claims that it is likely to rain on a particular day with probability $0.7$. The meteorologist would be considered calibrated (Dawid, 1982) when the following occurs: if $0.7$ is predicted on $D$ different days of the year, then it indeed rains on roughly $0.7D$ of those days (and the same holds for other probabilities). In the same sense, an ML classification model would be considered more reliable if it makes calibrated probabilistic predictions for the classes (Platt, 1999; Zadrozny and Elkan, 2001). A popular notion of calibration in multiclass classification settings is a simple 'one-dimensional' reduction to binary calibration proposed by Guo et al. (2017), called confidence calibration, which we recap below.

Let $\mathcal{X}$ and $[L] := \{1, 2, \ldots, L\}$ denote the feature and label spaces respectively. Consider a random point $(X, Y)$ drawn from some distribution $P$ over $\mathcal{X} \times [L]$. Let $c : \mathcal{X} \to [L]$ denote a classifier and $h : \mathcal{X} \to [0, 1]$ a function that provides a confidence score associated with the predicted class $c(X)$. The predictor $(c, h)$ is said to be confidence calibrated (for $P$) if $P(Y = c(X) \mid h(X)) = h(X)$. In other words, the fraction of instances where the predicted top label is correct, when the confidence $h(X)$ is $p \in [0, 1]$, approximately equals $p$. It is common to measure the confidence-miscalibration of $(c, h)$ using the expected-calibration-error (conf-ECE):

$$\text{conf-ECE}(c, h) := \mathbb{E}_X \left| P(Y = c(X) \mid h(X)) - h(X) \right|. \tag{1}$$

While confidence calibration is a reasonable minimum requirement, it is far from sufficient. Confidence calibration merges predictions with the same value of $h(X)$ across all classes; this merging can lead to predictors that have low conf-ECE but are not calibrated in any meaningful way.

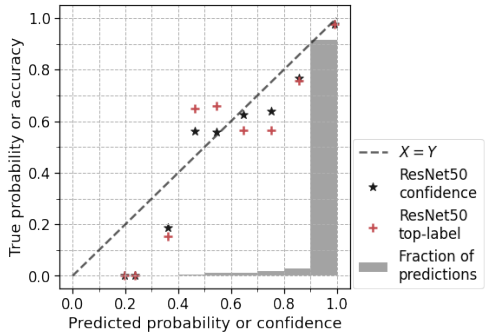
**Example 1** (Low conf-ECE does not imply meaningful calibration)**.** We construct a predictor which is clearly miscalibrated but has conf-ECE $= 0$. Suppose the feature space is $\mathcal{X} = \{a, b\}$ and $P(X = a) = P(X = b) = 0.5$. Consider a predictor $(c, h)$ with class predictions $c(a) = 1$, $c(b) = 2$, and confidence values $h(a) = h(b) = 0.6$. Let the true probabilities for the predicted class be $P(Y = 1 \mid X = a) = 0.2$ and $P(Y = 2 \mid X = b) = 1.0$. Observe that no matter what $X$ is, the predicted probability of the top-label is $0.6$, whereas the true probability is either $0.2$ or $1.0$. Despite this, the conf-ECE of $(c, h)$ is $0$ since $P(Y = c(X) \mid h(X) = 0.6) = 0.5\ P(Y = 1 \mid X = a) + 0.5\ P(Y = 2 \mid X = b) = 0.5(0.2 + 1) = 0.6$.

The reason for this strange behavior is that the probability $P(Y = c(X) \mid h(X))$ is not interpretable for decision-making. In practice, we always report both $c(X)$ and $h(X)$. Thus it is more reasonable to consider the conditional probability $P(Y = c(X) \mid c(X), h(X))$. Based on this understanding, we say that $(c, h)$ is *top-label calibrated* if

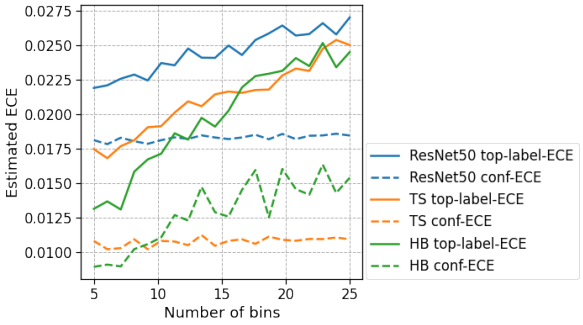
$$P(Y = c(X) \mid h(X), c(X)) = h(X).$$

[1]Carnegie Mellon University. Correspondence to: Chirag Gupta <chiragg@cmu.edu>.

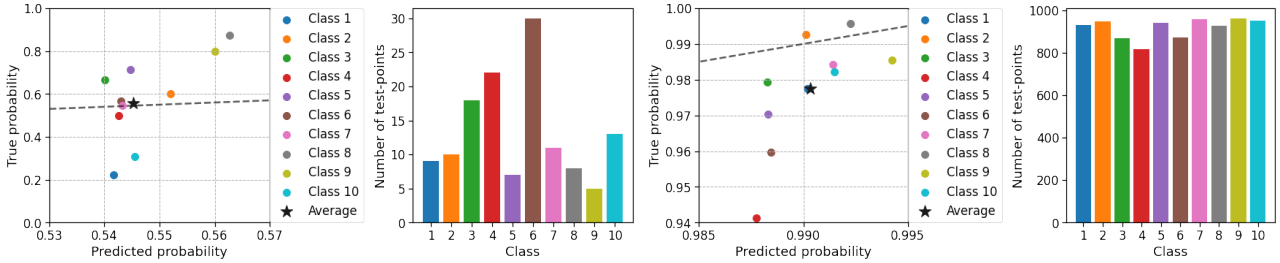
*Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.*, 




(a) Confidence reliability diagram (points marked ★) and top-label reliability diagram (points marked +) for a pre-temperature scaling ResNet50 model. The confidence reliability diagram (mistakenly) suggests better calibration, unlike the top-label reliability diagram.




(b) Conf-ECE and TL-ECE estimates of base ResNet50 model, temperature scaled (TS) model, and histogram binning (HB) model as the number of bins is varied. Binning is competitive with temperature scaling when assessing TL-ECE (green/orange solid curves), and appears to have lower TL-ECE when the number of bins is small.




(c) Class-wise and zoomed-in version of Figure 1a for bin 6 (left) and bin 10 (right). The markers ★ are in the same position as Figure 1a, and denote the average predicted and true probabilities. The colored points denote the predicted and true probabilities when seen class-wise. The histograms show the number of test points per class in bins 6 and 10.

*Figure 1.* Confidence reliability diagrams and conf-ECE scores underestimate the effective miscalibration of ResNet50 on CIFAR-10.

See the last paragraph of this section for a disambiguating remark on terminology. Define the top-label-ECE as

$$\text{TL-ECE}(c, h) := \mathbb{E}_X \left| P(Y = c(X) \mid c(X), h(X)) - h(X) \right|. \tag{2}$$

The predictor in Example 1 has TL-ECE$(c, h) = 0.4$, revealing its miscalibration. In Appendix E, we prove the following proposition.

**Proposition 1.** *For any* $(c, h)$, *conf-ECE*$(c, h) \leq$ *TL-ECE*$(c, h)$.

In practice, more benign versions of Example 1 occur, where conf-ECE is not completely meaningless, but can be misleading. Figure 1 illustrates this through the (test-time) performance of a ResNet50 model (He et al., 2016) on the CIFAR-10 dataset (Krizhevsky, 2009). The ★ markers in Figure 1a show a confidence reliability diagram, constructed using $B = 10$ fixed-width bins. The barplot indicates the fraction of $h(X)$ values that belong to each bin. The top-label reliability diagram assesses reliability when further conditioning on the predicted class. There are $L = 10$ different top-label reliability diagrams, so in order to make a comparison to confidence calibration, we estimate the average miscalibration across classes,

$$\Delta_b := \mathbb{E}\left[ \left| P(Y = c(X) \mid c(X), h(X)) - h(X) \right| \mid h(X) \in \text{Bin } b \right],$$

and plot $(\text{conf}_b, \text{conf}_b + \Delta_b)$ if $\text{acc}_b > \text{conf}_b$; $(\text{conf}_b, \text{conf}_b - \Delta_b)$ otherwise (these are the + markers in Figure 1a). We find that there is a visible increase in miscalibration when going from confidence calibration to top-label calibration. To understand why this change occurs, let us zoom into bin 6 ($h(X) \in [0.5, 0.6)$) and bin 10 ($h(X) \in [0.9, 1.0]$). Figure 1c displays the class-wise top-label reliability diagrams for these bins. Note that for bin 6, the ★ marker is nearly on the $X = Y$ line, indicating that the overall accuracy matches the overall confidence. However, the true accuracy when class 1 was predicted is $\approx 0.2$ and the true accuracy when class 8 was predicted is $\approx 0.9$ (a very similar scenario to Example 1). For bin 10, the ★ marker indicates a miscalibration of $\approx 0.01$; however, when class 4 was predicted (roughly 8% of all test-points) the miscalibration is $\approx 0.05$.

Figure 1b displays the aggregate effect of the above phenomenon (across bins and classes) through binned estimates of the conf-ECE and TL-ECE. This plot also displays the ECE estimates when the base model is recalibrated using temperature scaling (Guo et al., 2017) and histogram binning (Gupta and Ramdas, 2021; Zadrozny and Elkan, 2001). Since ECE estimates depend on the number of bins $B$, we varied $B$ in the range $[5, 25]$ to obtain unambigu-

ous conclusions. We find that the TL-ECE of the pre- and post- temperature scaling model is significantly higher than the conf-ECE across all values of $B$. For example, when $B = 15$, we observe that the TL-ECE of the temperature scaled ResNet model is nearly double its conf-ECE (going from $\approx 0.011$ to $\approx 0.022$). Further, while histogram binning appears worse than temperature scaling when assessing conf-ECE, it performs comparable or better than temperature scaling when assessing TL-ECE. Section 2 discusses top-label histogram binning in further detail.

**Post-hoc calibration setup.** This work is in the standard post-hoc calibration setup, where we wish to recalibrate an existing classifier using calibration data. The forthcoming Algorithm 1 assumes access to an existing class predictor $c : \mathcal{X} \to [L]$ and confidence predictor $g : \mathcal{X} \to [0, 1]$. The goal is to recalibrate $g$ to $h : \mathcal{X} \to [0, 1]$. Further details on this and relationships to other paradigms for multiclass calibration are discussed in Appendix A.

**Related works.** Issues with confidence calibration have been noticed before in other works (Kull et al., 2019; Vaicenavicius et al., 2019; Widmann et al., 2019), but the proposed solution in these works has been to consider notions of calibration that require more than just the top-label to be calibrated. Our work continues to advocate for calibrating probabilities only for the predicted class, which may be sufficient in many domains. Nevertheless, in Section 3 we develop histogram binning for class-wise calibration.

**Note on terminology.** The term conf-ECE was introduced by Kull et al. (2019). Most works refer to conf-ECE as just ECE (Guo et al., 2017; Kumar et al., 2018; Mukhoti et al., 2020; Nixon et al., 2020). However, some papers refer to conf-ECE as 'top-label-ECE' (Kumar et al., 2019; Zhang et al., 2020), resulting in two different terms for the same concept. We continue to call conf-ECE by its original name, and *our definition of top-label calibration and top-label ECE (2) is different from previous ones.*

## 2. Top-label histogram binning

Histogram binning (Gupta and Ramdas, 2021; Zadrozny and Elkan, 2001) is a post-hoc calibration method where the base model and calibration data is used to partition $\mathcal{X}$ into a number of 'bins', and the empirical distribution of the calibration $Y_i$ values in each bin is used to recalibrate $g$. Algorithm 1 describes an HB algirithm for top-label calibration. We start with a (miscalibrated) predictor $(c : \mathcal{X} \to [L], g : \mathcal{X} \to [0, 1])$. Then, the calibration data $\mathcal{D}$ is divided into $L$ different datasets $\{\mathcal{D}_l : l \in [L]\}$ based on the predicted class $c(X_i)$ for each point. Now for every $l \in [L]$, we calibrate $g$ to $h_l : \mathcal{X} \to [0, 1]$ using $\mathcal{D}_l$ and binary HB (line 4 calls Algorithm 2 of Gupta and Ramdas (2021)). The final predictor is $h(\cdot) = h_{c(\cdot)}(\cdot)$. The top-label predictor $c(\cdot)$ does not change in this process, and thus the accuracy remains the same.

**Algorithm 1:** Top-label histogram binning

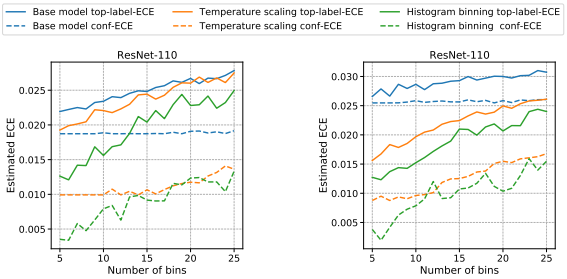
**Input:** Base predictor $(c, g)$, calibration data $\mathcal{D} = (X_1, Y_1), \ldots, (X_n, Y_n)$
**Hyperparameter:** # points per bin $k \in \mathbb{N}$ (say 50), tie-breaking parameter $\delta > 0$ (say $10^{-10}$)
**Output:** Top-label calibrated predictor $(c, h)$

1 **for** $l \leftarrow 1$ ***to*** $L$ **do**
2 $\quad \mathcal{D}_l \leftarrow \{(X_i, \mathbb{1}\{Y_i = l\}) : c(X_i) = l)\}$;
3 $\quad n_l \leftarrow |\mathcal{D}_l|$;
4 $\quad h_l \leftarrow$ Binary-HB$(g, \mathcal{D}_l, \lfloor n_l/k \rfloor, \delta)$;
5 **end**
6 $h(\cdot) \leftarrow h_{c(\cdot)}(\cdot)$;
7 **return** $(c, h)$;

While previous empirical works on HB fixed the *number of bins per class*, the analysis of Gupta and Ramdas (2021) suggests that a more principled way of choosing the number of bins is to fix the *number of points per bin*. This is parameter $k$ of Algorithm 1. Given $k$, the number of bins is decided separately for every class based on the amount of calibration data available. This is particularly relevant for top-label calibration since $n_l$, the number of points predicted as class $l$, can be non-uniform. Algorithm 1 adaptively sets the number of bins for class $l$ as $\lfloor n_l/k \rfloor$. The tie-breaking parameter $\delta$ can be arbitrarily small (like $10^{-10}$); it is used to ensure that outputs of different bins are not exactly identical by chance, so that conditioning on a calibrated probability output is equivalent to conditioning on a bin; this leads to a cleaner theoretical guarantee. In Appendix B, we discuss assumption-free theoretical guarantees for top-label HB, and experiments with a class-imbalanced dataset where fixing $k$ is practically useful.

Figures 2a and 2b compare the test-time performance of HB to temperature scaling (TS) on CIFAR-10 and CIFAR-100, with a ResNet-110 as the base model, and the loss function as Brier score or focal loss, which are typically better than cross-entropy for calibration (Mukhoti et al., 2020). Since there is no class imbalance in CIFAR-10, and to have unambiguous conclusions, we used a fixed number of bins per-class for HB (this is contradictory to Algorithm 1). We ranged the number of bins range from 5 to 25. The ECE of the base model and TS was estimated using fixed-width binning, with the same number of bins as HB.

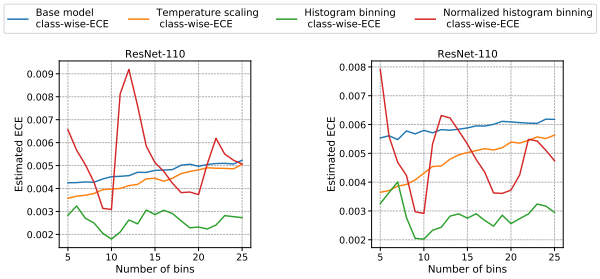
As previewed in Section 1, Figures 2a and 2b indicate that (a) the TL-ECE estimate of TS is significantly higher than the conf-ECE estimate, and (b) HB performs better than TS with respect to the TL-ECE estimates, across different numbers of bins. In Appendix D we define the max-




(a) CIFAR-10, focal loss. (b) CIFAR-10, Brier score.

(c) CIFAR-100, focal loss. (d) CIFAR-100, Brier score.

*Figure 2.* ECE of histogram binning (HB) and temperature scaling with ResNet-110 on CIFAR-10 and CIFAR-100. HB has lower estimated TL-ECE on CIFAR-10, but higher on CIFAR-100.




(a) CIFAR-10, focal loss. (b) CIFAR-10, Brier score.

(c) CIFAR-100, focal loss. (d) CIFAR-100, Brier score.

*Figure 3.* CW-ECE of non-normalized HB, normalized HB, and TS, with ResNet-110 on CIFAR-10 and CIFAR-100. Non-normalized HB performs the best overall.

imum calibration error metric (MCE) metric (Naeini et al., 2015) with respect to top-label calibration, and compare HB and TS based on the MCE. We find that the relative performance of HB for MCE is drastically better than TS.

CIFAR-100 has 100 classes and 5000 points for validation/calibration. Due to random subsampling, the validation split we used had one of the classes predicted as the top-label only 31 times. Thus very few points are available for recalibration when split across classes, and we do not expect HB to have small TL-ECE. This is confirmed by the plots in Figures 2c and 2d. HB has higher estimated TL-ECE than TS for most values of the number of bins. However in Section 3 we show that for class-wise calibration, HB performs better than TS on CIFAR-100. This is because in the class-wise setting, 5000 points are available for recalibration, which is sufficient for HB.

Further results with other deep net architectures are presented in Appendix D.

The deterioration in performance of HB when few calibration points are available has been observed in other works (Gupta and Ramdas, 2021; Kull et al., 2019; Niculescu-Mizil and Caruana, 2005). Comparing our results to previous empirical studies, we believe that if sufficiently many points are available for recalibration, or the number of classes is small, then HB performs quite well. To be more precise, we expect HB to be competitive if at least 200 points per class can be held out for recalibration, and the number of points per bin is at least $k \geq 20$.

## 3. Class-wise histogram binning

Class-wise calibration (Zadrozny and Elkan, 2002) requires the full $L$-dimensional prediction vector to be calibrated. A predictor $\mathbf{h} = (h_1, h_2, \ldots, h_L)$ is said to be class-wise calibrated if for every $l \in [L]$, $P(Y = l \mid h_l(X)) = h_l(X)$. A measure of the class-wise calibration of $(c, \mathbf{h})$ is the class-wise-ECE (Kull et al., 2019):

$$\text{CW-ECE}(c, \mathbf{h}) := L^{-1} \sum_{l=1}^{L} \mathbb{E}_X \left| P(Y = l \mid h_l(X)) - h_l(X) \right| . \tag{3}$$

A common way to achieve class-wise calibration is to use a binary calibration algorithm in a 1-v-all paradigm, to learn $L$ predictors $h_1, h_2, \ldots, h_L$, which are then normalized so that the final prediction sums to one. While the normalization makes sense for interpretability, it is not required by the class-wise calibration condition. In Appendix C, we formally state a *non-normalized* class-wise HB algorithm and show calibration guarantees for it. We were unable to derive such guarantees for normalized HB. We hypothesized that normalization could hurt the performance of HB in practice as well, and found that this is indeed the case. Figure 3 presents estimates of the CW-ECE for CIFAR-10 and CIFAR-100 with ResNet-110 (results with other architectures are presented in Appendix C). Non-normalized 1-v-all HB performs better than TS in all our experiments, and normalized HB performs worse than even the base model. Guo et al. (2017) and Kull et al. (2019) have also noted such negative results for normalized HB. Our experiments reveal that these negative results are not inherent to HB but are simply due to an incorrect normalization step.

## References


Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24 (3):131–151, 1999.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, 2021.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, 2020.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2018.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 2005.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2020.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: a unifying framework. In *Advances in Neural Information Processing Systems*, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning*, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, 2002.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

Jize Zhang, Bhavya Kailkhura, and T Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 2020.

| | $g(a)$ | $g(b)$ | $g(c)$ | $g(d)$ | $g(e)$ | $g(f)$ |
|---|---|---|---|---|---|---|
| $g_1(\cdot)$ | 0.1 | 0.6 | 0.2 | 0.0 | 0.0 | 0.9 |
| $g_2(\cdot)$ | 0.0 | 0.0 | 0.7 | 0.1 | 0.1 | 0.1 |
| $g_3(\cdot)$ | 0.6 | 0.1 | 0.0 | 0.1 | 0.8 | 0.0 |
| $g_4(\cdot)$ | 0.3 | 0.3 | 0.1 | 0.8 | 0.1 | 0.0 |
| $g(\cdot)$ | 0.6 | 0.6 | 0.7 | 0.8 | 0.8 | 0.9 |
| $c(\cdot)$ | 3 | 1 | 2 | 4 | 3 | 1 |
| $Y$ | 3 | 4 | 2 | 1 | 4 | 1 |

(a) Predictions of a fixed base model $\mathbf{g} : \mathcal{X} \to \Delta^3$ on calibration data $\mathcal{D} = \{(a, 3), (b, 4), \ldots, (f, 1)\}$. Note for each $x$, $\sum_{l=1}^{4} g_l(x) = 1$.

| $g(\cdot)$ | 0.6 | 0.6 | 0.7 | 0.8 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|
| $Y = c(\cdot)$ | 1 | 0 | 1 | 0 | 0 | 1 |

(b) Confidence calibration

| $c(\cdot) = 1$ | $g(\cdot)$ | 0.6 | 0.9 |
|---|---|---|---|
| | $Y = c(\cdot)$ | 0 | 1 |

| $c(\cdot) = 2$ | $g(\cdot)$ | 0.7 |
|---|---|---|
| | $Y = c(\cdot)$ | 1 |

| $c(\cdot) = 3$ | $g(\cdot)$ | 0.6 | 0.8 |
|---|---|---|---|
| | $Y = c(\cdot)$ | 1 | 0 |

| $c(\cdot) = 4$ | $g(\cdot)$ | 0.8 |
|---|---|---|
| | $Y = c(\cdot)$ | 0 |

(c) Top-label calibration

| $g_1(\cdot)$ | 0.1 | 0.6 | 0.2 | 0.0 | 0.0 | 0.9 |
|---|---|---|---|---|---|---|
| $Y = 1$ | 0 | 0 | 0 | 1 | 0 | 1 |

| $g_2(\cdot)$ | 0.0 | 0.0 | 0.7 | 0.1 | 0.1 | 0.1 |
|---|---|---|---|---|---|---|
| $Y = 2$ | 0 | 0 | 1 | 0 | 0 | 0 |

| $g_3(\cdot)$ | 0.6 | 0.1 | 0.0 | 0.1 | 0.8 | 0.0 |
|---|---|---|---|---|---|---|
| $Y = 3$ | 1 | 0 | 0 | 0 | 0 | 0 |

| $g_4(\cdot)$ | 0.3 | 0.3 | 0.1 | 0.8 | 0.1 | 0.0 |
|---|---|---|---|---|---|---|
| $Y = 4$ | 0 | 1 | 0 | 0 | 1 | 0 |

(d) Class-wise calibration

| $g(a)$ | $Y(a)$ | $g(b)$ | $Y(b)$ | ... | $g(f)$ | $Y(f)$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.0 | 0.2 | 0.0 | .. | 0.9 | 1.0 |
| 0.0 | 0.0 | 0.7 | 0.0 | .. | 0.1 | 0.0 |
| 0.6 | 1.0 | 0.0 | 0.0 | .. | 0.0 | 0.0 |
| 0.3 | 0.0 | 0.1 | 1.0 | .. | 0.0 | 0.0 |

(e) Canonical calibration

*Figure 4.* Paradigms of post-hoc multiclass calibration from most lenient (b) to most stringent (e). Plots (b–e) show the input for a calibration method that aims to achieve the corresponding notion. The numbers in these tables are derived from a fixed base classifier $\mathbf{g}$ applied to the calibration data $\mathcal{D}$, given in plot (a). The top label is $c(\cdot)$, and $g(\cdot)$ is its confidence. Appendix A has further details.

## A. Post-hoc calibration setting

This paper considers the standard recalibration or post-hoc calibration setting. We start with a fixed 'pre-learnt' base model $\mathbf{g} : \mathcal{X} \to \Delta^{L-1}$, where $\mathcal{X}$ denotes the feature space and $\Delta^{L-1}$ is the probability simplex in $L$ dimensions. (To ease readability, we typically use boldface characters such as $\mathbf{g}, \mathbf{h}, \mathbf{Y}$ to denote elements of $\Delta^{L-1}$ or functions whose range is $\Delta^{L-1}$.) The base model $\mathbf{g}$ can correspond to a deep net, a random forest, or any 1-v-all (one-versus-all) binary classification model such as logistic regression. The base model is typically optimized for classification accuracy and may not be calibrated. The goal of post-hoc calibration is to use some given *calibration data* $\mathcal{D} = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in (\mathcal{X} \times [L])^n$, typically data on which $\mathbf{g}$ was not learnt, to recalibrate $\mathbf{g}$. We use $(X, Y)$ to denote a random point identically distributed as the test point. For theoretical analysis, we assume that $\mathcal{D}$ is a fresh sample statistically independent of $\mathbf{g}$, and the $(X_i, Y_i)$'s are independent and identically distributed as $(X, Y)$.

*Our theoretical guarantees make no assumptions other than this i.i.d. assumption.* Thus our guarantees hold even when the data-generating distribution is arbitrarily non-smooth. This is known as the distribution-free setting (Gupta et al., 2020). In practice, $\mathcal{D}$ is often the same as the validation data; if $\mathcal{D}$ is used for hyperparameter tuning, early stopping, etc., then the i.i.d. assumption is violated.

Post-hoc calibration for multiclass classification can occur under four different paradigms, illustrated in Figure 4. Letting $g_l$ denote the $l$'th component of $\mathbf{g}$, we can derive a top-label class predictor $c : \mathcal{X} \to [L]$ and a top-label confidence predictor $g : \mathcal{X} \to [0, 1]$ given by

$$c(\cdot) = \operatorname*{arg\,max}_{l \in [L]} \mathbf{g}_l(\cdot), \quad \text{and} \quad g(\cdot) = \max_{l \in [L]} \mathbf{g}_l(\cdot), \tag{4}$$

assuming an arbitrary tie-breaking rule for the $\arg\max$. Confidence calibration (Figure 4b) cares only about $g$ and the indicators $\mathbb{1}\{Y_i = c(X_i)\}$. Top-label calibration (Figure 4c) also focuses on $g$, but additionally requires access to the actual label predictions $c(X_i)$. This is used to produce one predictor for each label, denoted as $h_1, h_2, \ldots, h_L$, but the final predictor for top-label calibration is a single $h$ that predicts the $h_l$ corresponding to the top-label: $h = h_{c(\cdot)}(\cdot)$. The class predictor $c$ is not changed in this process, and thus the accuracy of $(c, h)$ is the same as the accuracy of $(c, g)$. Class-wise calibration (Figure 4d, formally defined and discussed in Section 3) also produces $h_1, h_2, \ldots, h_L$, but $h_l$ is required to be calibrated whether or not $l$ is the top-label. Unlike top-label calibration, a 1-v-all class-wise calibration method would use the $g_l$ values on all $n$ calibration points when learning $h_l$.

## B. Distribution-free guarantees for top-label histogram binning

In this section, we show distribution-free calibration guarantees for top-label HB (Algorithm 1). Top-label HB recalibrates $g$ to a piecewise constant function $h$ that takes one value per bin. Consider a specific bin $b$; the $h$ value for this bin is computed as the average of the indicators $\{\mathbb{1}\{Y_i = c(X_i)\} : X_i \in \text{Bin } b\}$. This is an estimate of the 'bias' of the bin $P(Y = c(X) \mid X \in \text{Bin } b)$. A concentration inequality can then be used to bound the deviation between the estimate and the true bias to prove distribution-free calibration guarantees. In the forthcoming Theorem 1, we show high-probability and in-expectation bounds on the the TL-ECE of the predictor learnt by HB. Additionally, we show marginal and conditional top-label calibration bounds. These notions were proposed in the binary calibration setting by Gupta et al. (2020) and Gupta and Ramdas (2021). In the definition below, $\mathcal{A}$ refers to any algorithm that takes as input calibration data $\mathcal{D}$ and an initial classifier $\mathbf{g}$ to produce a top-label predictor $c$ and an associated probability map $h$. Algorithm 1 is an example of $\mathcal{A}$.

**Definition 1** (Marginal and conditional top-label calibration)**.** Let $\varepsilon, \alpha \in (0,1)$ be some given levels of approximation and failure respectively. An algorithm $\mathcal{A} : (\mathbf{g}, \mathcal{D}) \mapsto (c, h)$ is

(a) $(\varepsilon, \alpha)$-marginally top-label calibrated if for every distribution $P$ over $\mathcal{X} \times [L]$,

$$P\Big(|P(Y = c(X) \mid c(X), h(X)) - h(X)| \le \varepsilon\Big) \ge 1 - \alpha. \tag{5}$$

(b) $(\varepsilon, \alpha)$-conditionally top-label calibrated if for every distribution $P$ over $\mathcal{X} \times [L]$,

$$P\Big(\forall\, l \in [L], r \in \text{Range}(h), |P(Y = c(X) \mid c(X) = l, h(X) = r) - r| \le \varepsilon\Big) \ge 1 - \alpha. \tag{6}$$

To clarify, all probabilities are taken over the randomness in the new point $(X, Y)$ and in $\mathcal{D}$ (which is implicit in $c, h$ since they are produced by $\mathcal{A}(\mathcal{D}, \mathbf{g})$), all of which are i.i.d. from $P$. Marginal calibration asserts that with high probability, on average over the distribution of $\mathcal{D}, X$, $P(Y = c(X) \mid c(X), h(X))$ is at most $\varepsilon$ away from $h(X)$. In comparison, TL-ECE is the average of these deviations over $X$. Marginal calibration may be a more appropriate metric for calibration than TL-ECE if we are somewhat agnostic to probabilistic errors less than some fixed threshold $\varepsilon$ (like 0.05). Conditional calibration is a strictly stronger definition that requires the deviation to be at most $\varepsilon$ for every possible prediction $(l, r)$, including rare ones, not just on average over predictions. This may be relevant in medical settings where we want the prediction on every patient to be reasonably calibrated. Algorithm 1 satisfies the following calibration guarantees.

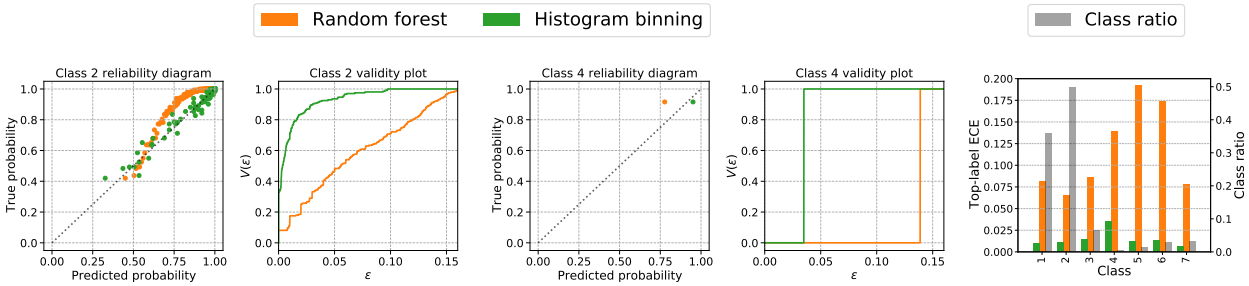
**Theorem 1.** *Fix hyperparameters $\delta > 0$ (arbitrarily small) and points per bin $k \ge 2$, and assume $n_l \ge k$ for every $l \in [L]$. Then, for any $\alpha \in (0,1)$, Algorithm 1 is $(\varepsilon_1, \alpha)$-marginally and $(\varepsilon_2, \alpha)$-conditionally top-label calibrated:*

$$\varepsilon_1 = \sqrt{\frac{\log(2/\alpha)}{2(k-1)}} + \delta, \qquad \textit{and} \qquad \varepsilon_2 = \sqrt{\frac{\log(2n/k\alpha)}{2(k-1)}} + \delta. \tag{7}$$

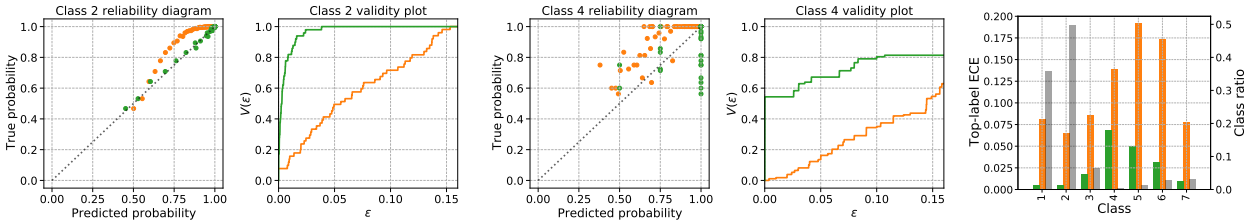
*Further, $P(\text{TL-ECE}(c, h) \le \varepsilon_2) \ge 1 - \alpha$ , and $\mathbb{E}\left[\text{TL-ECE}(c, h)\right] \le \sqrt{1/2k} + \delta$.*

The proof in Appendix E is a multiclass top-label adaption of the guarantee in the binary setting by Gupta and Ramdas (2021). The $\widetilde{O}(1/\sqrt{k})$ dependence of the bound relies on Algorithm 1 delegating at least $k$ points to every bin. Since $\delta$ can be chosen to be arbitrarily small, setting $k = 50$ gives roughly $\mathbb{E}_{\mathcal{D}}\left[\text{TL-ECE}(h)\right] \le 0.1$. Base on this, we suggest setting $k \in [50, 150]$ in practice.

The principled methodology of fixing the number of points per bin reaps practical benefits. Figure 5 illustrates this through the performance of HB for the class imbalanced COVTYPE-7 dataset (Blackard and Dean, 1999) with class ratio approximately 36% for class 1 and 49% for class 2. The entire dataset has 581012 points which is divided into train-test in the ratio 70:30. Then, 10% of the training points are held out for calibration ($n = |\mathcal{D}| = 40671$). The base classifier is a random forest (RF) trained on the remaining training points (it achieves around 95% test accuracy). The RF is then recalibrated using HB. The top-label reliability diagrams in Figure 5a illustrate that the original RF (in orange) is *underconfident* on both the most likely and least likely classes. Additional figures in Appendix D show that the RF is always underconfident no matter which class is predicted as the top-label. HB (in green) recalibrates the RF effectively across all classes. Validity plots (Gupta and Ramdas, 2021) estimate how the LHS of condition (5), denoted as $V(\varepsilon)$, varies with $\varepsilon$. We observe that for all $\varepsilon$, $V(\varepsilon)$ is higher for HB. The rightmost barplot compares the estimated TL-ECE for all classes, and also shows the class proportions. While the original RF is significantly miscalibrated for the less likely classes, HB has a more uniform miscalibration across classes. Figure 5b considers a slightly different HB algorithm where the number of points per class is not adapted to the number of times the class is predicted, but is fixed beforehand (this corresponds to replacing $\lfloor n_l/k \rfloor$ in line 4 of Algorithm 1 with a fixed $B \in \mathbb{N}$). While even in this setting there is a drop in the TL-ECE compared to the RF model, the final profile is less uniform compared to fixing the number of points per bin.




(a) Top-label histogram binning (Algorithm 1) with $k = 100$ points per bin. Class 4 has only 183 calibration points. Algorithm 1 adapts and uses only a single bin to ensure that the TL-ECE on Class 4 is comparable to the TL-ECE on Class 2. Overall, the random forest classifier has significantly higher TL-ECE for the least likely classes (4, 5, and 6), but the post-calibration TL-ECE using binning is quite uniform.




(b) Histogram binning with $B = 50$ bins for every class. Compared to Figure 5a, the post-calibration TL-ECE for the most likely classes decreases while the TL-ECE for the least likely classes increases.

*Figure 5.* Recalibration of a random forest using histogram binning on the class imbalanced COVTYPE-7 dataset (Class 2 is roughly 100 times likelier than Class 4). By ensuring a fixed number of calibration points per bin, Algorithm 1 obtains relatively uniform top-label calibration across classes (Figure 5a). In comparison, if a fixed number of bins are chosen for all classes, the performance deteriorates for the least likely classes (Figure 5b).

## C. Distribution-free class-wise calibration using histogram binning

We define marginal and conditional calibration for class-wise calibration, analogous to Definition 1, and state a histogram binning algorithm that is calibrated with respect to these notions. We also show bounds on the CW-ECE of the proposed algorithm (CW-ECE is defined in equation (3)).

A general algorithm $\mathcal{A}$ for class-wise calibration takes as input calibration data $\mathcal{D}$ and an initial classifier $\mathbf{g}$ to produce an approximately class-wise calibrated predictor $\mathbf{h} : \mathcal{X} \to [0,1]^L$. Recall that we denote the component functions of $\mathbf{h}$ as $h_1, h_2, \ldots, h_L$. In the same fashion, define the notation $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L) \in (0,1)^L$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_L) \in (0,1)^L$.

**Definition 2** (Marginal and conditional class-wise calibration)**.** Let $\boldsymbol{\varepsilon}, \boldsymbol{\alpha} \in (0,1)^L$ be some given levels of approximation and failure respectively. An algorithm $\mathcal{A} : (\mathbf{g}, \mathcal{D}) \mapsto \mathbf{h}$ is

(a) $(\boldsymbol{\varepsilon}, \boldsymbol{\alpha})$-marginally class-wise calibrated if for every distribution $P$ over $\mathcal{X} \times [L]$ and for every $l \in [L]$

$$P\Big( |P(Y = l \mid h_l(X)) - h_l(X)| \le \varepsilon_l \Big) \ge 1 - \alpha_l. \tag{8}$$

(b) $(\boldsymbol{\varepsilon}, \boldsymbol{\alpha})$-conditionally class-wise calibrated if for every distribution $P$ over $\mathcal{X} \times [L]$ and for every $l \in [L]$,

$$P\Big(\forall r \in \text{Range}(h), |P(Y = l \mid h_l(X) = r) - r| \le \varepsilon_l \Big) \ge 1 - \alpha_l. \tag{9}$$

Definition 2 requires that each $h_l$ is $(\varepsilon_l, \alpha_l)$ calibrated in the binary senses defined by Gupta et al. (2021, Definitions 1 and 2). Let $\alpha = \sum_{l=1}^{L} \alpha_l$ and $\varepsilon = \max_{l \in [L]} \varepsilon_l$. Then (8) implies

$$P\Big(\forall l \in [L], |P(Y = l \mid h_l(X)) - h_l(X)| \le \varepsilon \Big) \ge 1 - \alpha, \tag{10}$$

**Algorithm 2:** Class-wise histogram binning

**Input:** Base multiclass predictor $\mathbf{g}$, calibration data $\mathcal{D} = (X_1, Y_1), \ldots, (X_n, Y_n)$
**Hyperparameter:** # points per bin $k_1, k_2, \ldots, k_l \in \mathbb{N}^L$ (say each $k_l = 50$), tie-breaking parameter $\delta > 0$ (say $10^{-10}$)
**Output:** $L$ class-wise calibrated predictors $h_1, h_2, \ldots, h_L$

**for** $l \leftarrow 1$ ***to*** $L$ **do**
  $\mathcal{D}_l \leftarrow \{(X_i, \mathbb{1}\{Y_i = l\}) : i \in [n]\}$;
  $h_l \leftarrow$ Binary-histogram-binning$(g, \mathcal{D}_l, \lfloor n/k_l \rfloor, \delta)$;
**end**
**return** $(h_1, h_2, \ldots, h_L)$;

and (9) implies

$$P\Big(\forall l \in [L], r \in \text{Range}(h), |P(Y = l \mid h_l(X) = r) - r| \leq \varepsilon\Big) \geq 1 - \alpha. \tag{11}$$

The choice of not including the uniformity over $L$ in Definition 2 reveals the nature of our class-wise HB algorithm and the upcoming theoretical guarantees: (a) we learn the $h_l$'s separately for each $l$ and do not combine the learnt functions in any way such as normalization, (b) we do not combine the calibration inequalities for different $[L]$ in any other way other than a union bound. Thus the only way we can show (10) (or (11)) is by using a union bound over (8) (or (9)).

To achieve class-wise calibration using binary routines, we learn each component function $h_l$ in a 1-v-all fashion. Algorithm 2 contains the pseudocode with the underlying routine as binary HB. To learn $h_l$, we use a dataset $\mathcal{D}_l$, which unlike top-label HB (Algorithm 1), contains $X_i$ even if $c(X_i) \neq l$. However the $Y_i$ is replaced with $\mathbb{1}\{Y_i = l\}$. The number of points per bin $k_l$ can be different for different classes, but generally one would set $k_1 = \ldots = k_L = k \in \mathbb{N}$. Larger values of $k_l$ will lead to smaller $\varepsilon_l$ and $\delta_l$ in the guarantees, at loss of sharpness since the number of bins $\lfloor n/k_l \rfloor$ would be smaller.

We now state the distribution-free calibration guarantees satisfied by Algorithm 2.

**Theorem 2.** *Fix hyperparameters $\delta > 0$ (arbitrarily small) and points per bin $k_1, k_2, \ldots, k_l \geq 2$, and assume $n_l \geq k_l$ for every $l \in [L]$. Then, for every $l \in [L]$, for any $\alpha_l \in (0, 1)$, Algorithm 2 is $(\boldsymbol{\varepsilon}^{(1)}, \boldsymbol{\alpha})$-marginally and $(\boldsymbol{\varepsilon}^{(2)}, \boldsymbol{\alpha})$-conditionally class-wise calibrated with*

$$\varepsilon_l^{(1)} = \sqrt{\frac{\log(2/\alpha_l)}{2(k_l - 1)}} + \delta, \qquad \text{and} \qquad \varepsilon_l^{(2)} = \sqrt{\frac{\log(2n/k_l\alpha_l)}{2(k_l - 1)}} + \delta. \tag{12}$$

*Further,*

*(a)* $P(\textit{CW-ECE}(c, h) \leq \max_{l \in [L]} \varepsilon_l^{(2)}) \geq 1 - \sum_{l \in [L]} \alpha_l$, *and*

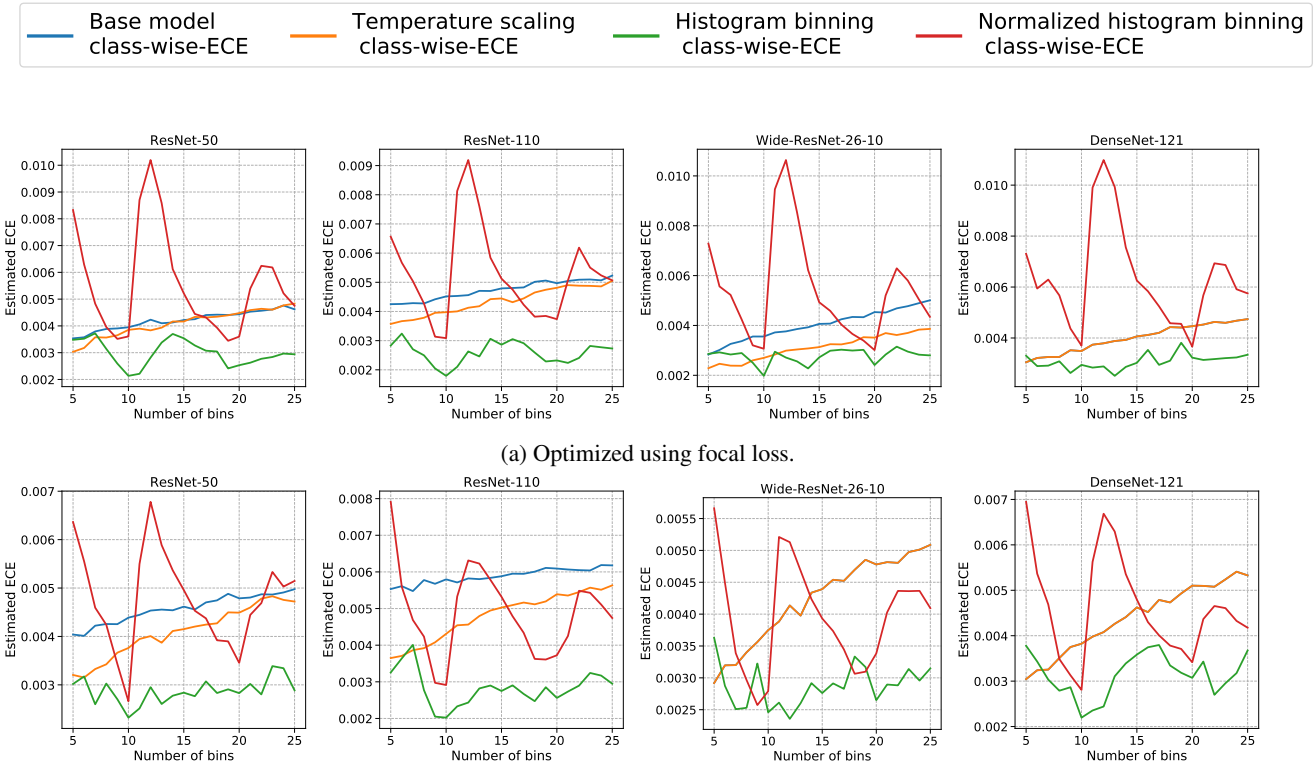
*(b)* $\mathbb{E}\left[\textit{CW-ECE}(c, h)\right] \leq \max_{l \in [L]} \sqrt{1/2k_l} + \delta$.

Theorem 2 is proved in Appendix E. The proof follows by using the result of Gupta and Ramdas (2021, Theorem 2), derived in the binary calibration setting, for each $h_l$ separately. As discussed in Section 3, unlike previous works (Guo et al., 2017; Kull et al., 2019; Zadrozny and Elkan, 2002), Algorithm 2 does not normalize the $h_l$'s. We do not know how to derive Theorem 2 style results for a normalized version of Algorithm 2.

Figure 6 presents estimated CW-ECE values for four deep net architectures trained using two loss functions on CIFAR-10 across different values for the numbers of bins, and Figure 7 presents the same results on CIFAR-100. These plots compare the base model, temperature scaling, non-normalized HB and normalized HB, same as the plots in Figure 3 of the main paper; the plots from the main paper are also reproduced to ease comparison. In both plots, we find that non-normalized HB performs much better than both the base model and temperature scaling, across all deep net architectures that we considered.

## D. Experimental details and additional results

This section presents further details and results to supplement those presented in Sections 1 and 2 of the main paper. Appendix D.1 contains discusses the COVTYPE-7 dataset. Appendix D.2 contains discusses the CIFAR datasets. In




(a) Optimized using focal loss.

(b) Optimized using Brier score.

*Figure 6.* CW-ECE of histogram binning (HB) and temperature scaling (TS) with deep nets on CIFAR-10. TS does not change the base model CW-ECE much; HB reduces it significantly, and normalization appears to hurt HB significantly.

Appendix D.2, we also define a top-label version of the maximum calibration error (MCE) metric (Guo et al., 2017; Naeini et al., 2015), and show results on the CIFAR datasets for this metric. Overall, we find that top-label HB typically performs better than temperature scaling for MCE.

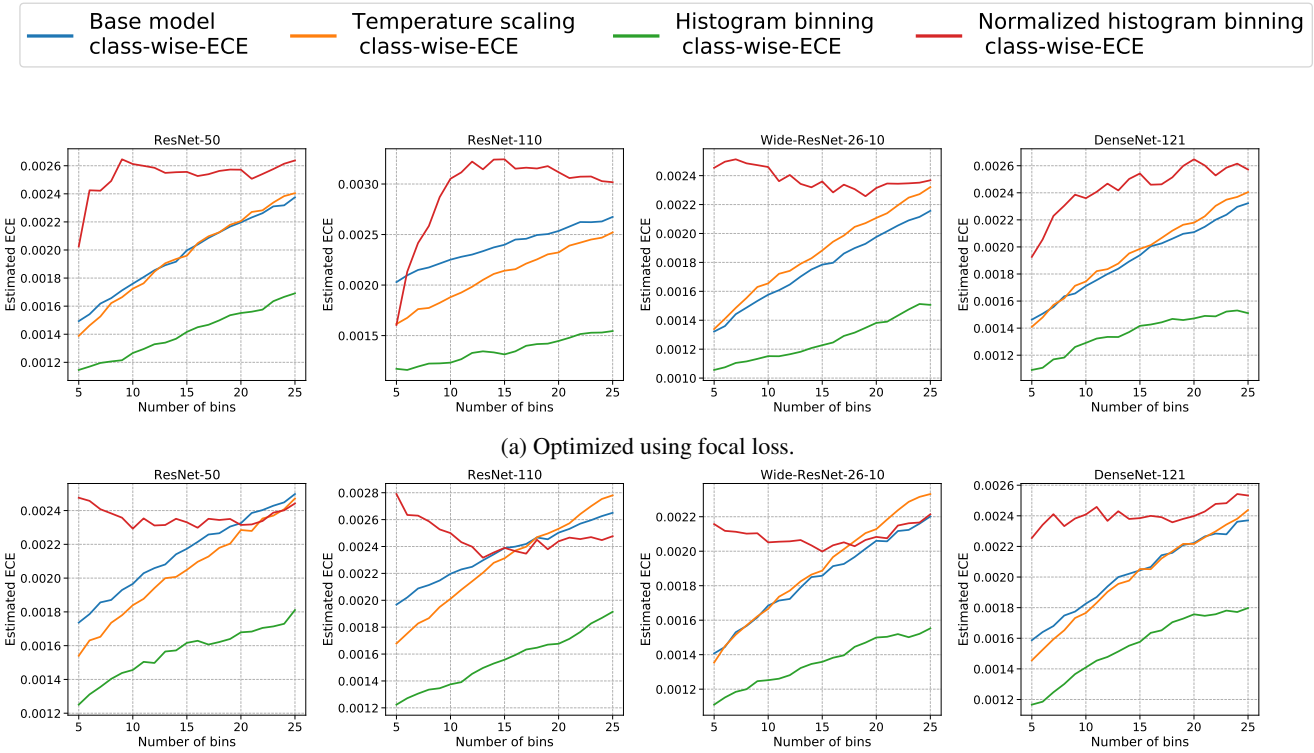
### D.1. Top-label calibration of COVTYPE-7

We present additional details and results for the top-label HB experiments of Section 2 and Appendix B. The base classifier is an RF learnt using `sklearn.ensemble import RandomForestClassifier` with default parameters. The base RF is a nearly continuous base model since most predictions are unique. Thus, we need to use binning to make reliability diagrams, validity plots, and perform ECE estimation, for the base model. To have a fair comparison, instead of having a fixed binning scheme to assess the base model, the binning scheme was decided based on the unique predictions of top-label HB. Thus for every $l$, and $r \in \text{Range}(h_l)$, the bins are defined as $\{x : c(x) = l, h_l(x) = r\}$. Thus while the base model in Figures 5a and 5b are the same, the reliability diagrams and validity plots in orange are different. As can be seen in the bar plots in Figure 5, the ECE estimation is not affected significantly.

When $k = 100$, the total number of bins that were chosen by Algorithm 1 was 403, which is roughly $57.6$ bins per class. The choice of $B = 50$ for the fixed bins per class experiment was made on this basis.

Figure 8 supplements Figure 5 in the main paper by presenting reliability diagrams and validity plots of top-label HB for all classes. Figure 8a presents the plots with adaptive number of bins per class (Algorithm 1), and Figure 8b presents these for fixed number of bins per class. We make the following observations.

(a) For every class $l \in [L]$, the RF is overconfident. This may seem surprising at first since we generally expect that models may be overconfidence for certain classes and underconfident for others. However, note that all our plots assess top-label calibration, that is, we are assessing the predicted and true probabilities of only the predicted class. It is possible that a model is overconfident for every class whenever that class is predicted to be the top-label.




(a) Optimized using focal loss.

(b) Optimized using Brier score.

*Figure 7.* CW-ECE of histogram binning (HB) and temperature scaling (TS) with deep nets on CIFAR-100. TS does not change the base model CW-ECE much; HB reduces it significantly, and normalization appears to hurt HB significantly.

(b) For the most likely classes, namely classes 1 and 2, the number of bins in the adaptive case is higher than 50. Fewer bins leads to better calibration (at the cost of sharpness). This can be verified through the validity plots for classes 1 and 2 — the validity plots in the fixed bins case is slightly 'above' the validity plot in the adaptive bin case. However both validity curves are quite similar.

(c) The opposite is true for the least likely classes, namely classes 4, 5, 6. The validity plot in the fixed bins case is 'below' the validity plot in the adaptive bins case, indicating higher TL-ECE in the fixed bins case. The difference between the validity plots is high. Thus if a fixed number of bins per class is pre-decided, the performance for the least likely classes significantly suffers.

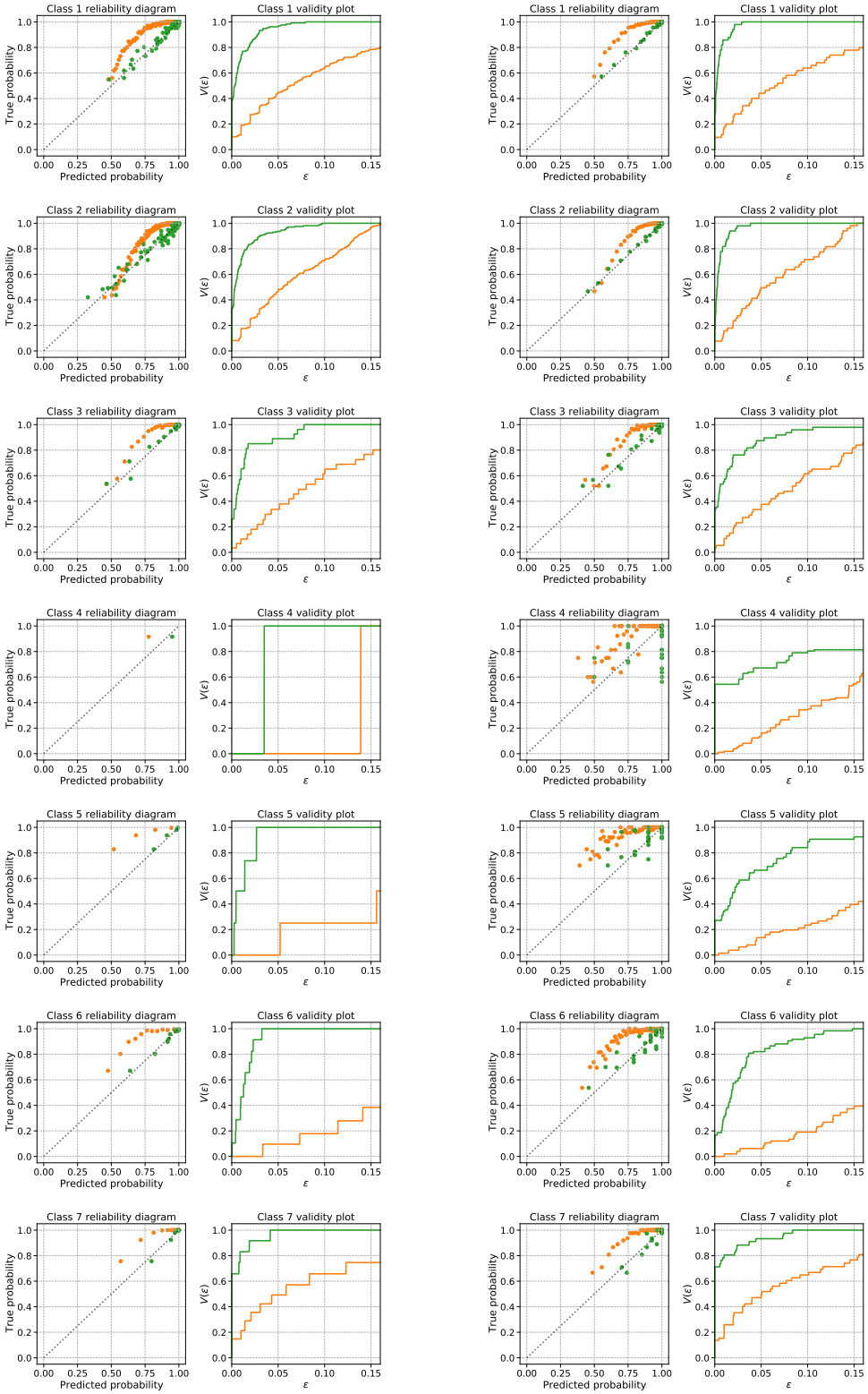
Based on these observations, we recommend adaptively choosing the number of bins per class, as done by Algorithm 1.

### D.2. Top-label calibration of CIFAR-10 and CIFAR-100

We present additional details and results for the top-label HB experiment in the main paper. We used the base models trained by Mukhoti et al. (2020) on the following architectures: ResNet-50, Resnet-110 (He et al., 2016), Wide-ResNet-26-10 (Zagoruyko and Komodakis, 2016), and DensetNet-121 (Huang et al., 2017).[1] Mukhoti et al. (2020) noted that using the loss function as the Brier score, or their proposed focal loss, leads to the best performance pre- and post- TS. Our experiment uses the exact base models that lead to the numbers in Table 1 of their paper (corresponding to the columns 'Brier Loss' and 'FLSD-53'), and the same calibration data that was used for TS was also used for HB. We did not perform any additional tuning, and our results are thus relatively free of selection biases.

The conf-ECE and TL-ECE estimation for the base model and temperature scaling was done using fixed-width bins $[0, 1/B), [1/B, 2/B), \ldots [1 - 1/B, 1]$. Plugin estimates of the ECE were used, same as Guo et al. (2017). The TL-

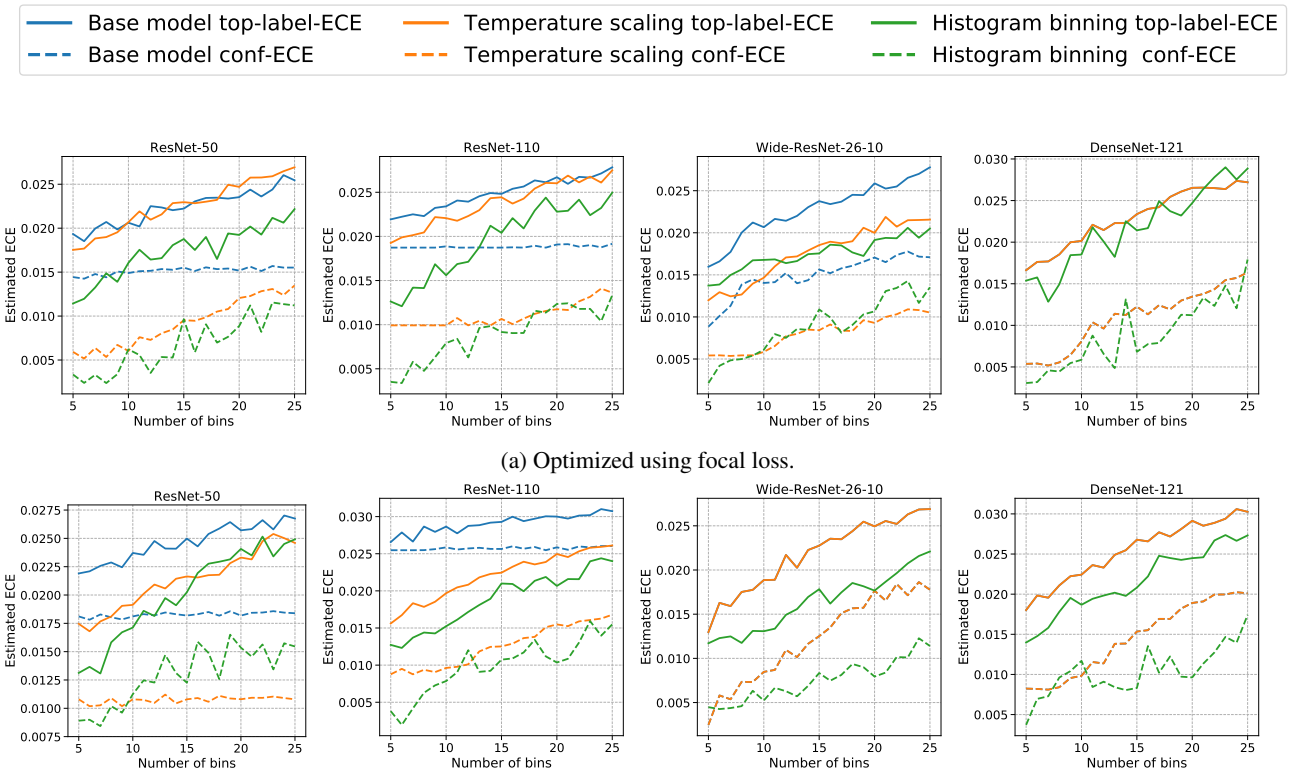
[1]The models were obtained from `www.robots.ox.ac.uk/~viveka/focal_calibration/` and used along with the code at `github.com/torrvision/focal_calibration` to obtain base predictions.




(a) Top-label HB with $k = 100$ points per bin.

(b) Top-label HB with $B = 50$ bins per class.

*Figure 8.* Top-label histogram binning (HB) calibrates a miscalibrated random-forest on the class imbalanced COVTYPE-7 dataset. For the less likely classes (4, 5, and 6), the left column is better calibrated than the right column. Similar observations are made on other datasets, and so we recommend adaptively choosing a different number of bins per class, as Algorithm 1 does.




(a) Optimized using focal loss.

(b) Optimized using Brier score.

*Figure 9.* Conf-ECE and TL-ECE of histogram binning and temperature scaling with deep nets on CIFAR-10. When assessed with respect to TL-ECE (solid lines in all plots), histogram binning is always comparable to temperature scaling, and often performs better.

ECE estimation for top-label HB did not use further binning since HB is already a discrete output method (due to Jensen's inequality, any further binning will only decrease the ECE estimate). The conf-ECE estimation for top-label HB is trickier since Range$(h_l)$ is different for different $l$, and thus the averaging across bins cannot be done unless we further define bins based on fixed-width binning. We observed in our experiments that fixed-width binning decreased the conf-ECE of top-label HB. To have a fairer comparison, instead of fixed-width binning, we merged the $k$'th bins across all classes for estimating conf-ECE. That is, for a given $k$, we considered the bins across classes corresponding to the $k/B$ and $(k+1)/B$ quantiles on calibration data, and computed the average confidence and accuracy for the test points that belong to these bins (across all predicted top-labels).

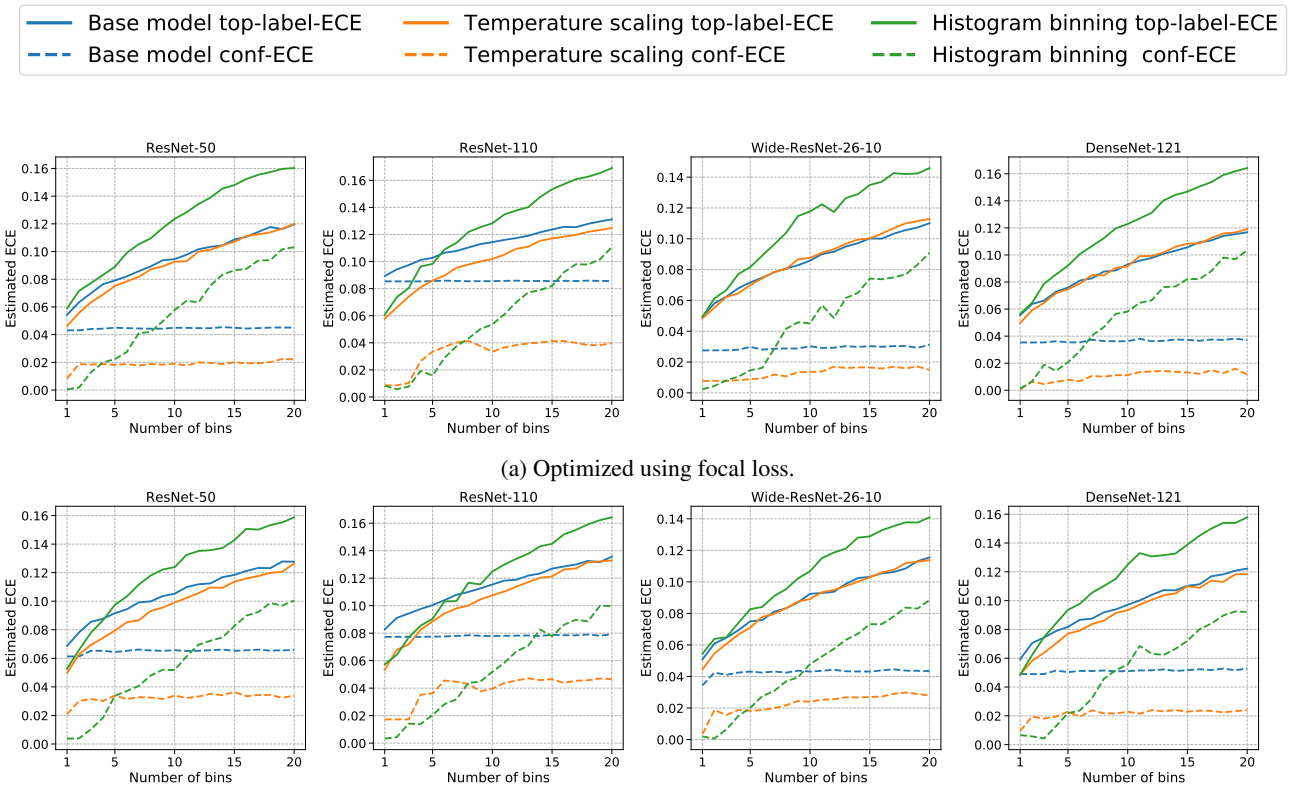
Figures 9 presents results of our experiments on additional deep net architectures with CIFAR-10, to supplement Figure 2 in the main paper. Figure 10 presents results with CIFAR-100. The observations are very similar to those presented in the main paper.

In some plots, such as the Wide-Resnet-26-10 plot in Figure 2b and Figure 9b, the blue base model line and the orange temperature scaling line intersect. This is not surprising; it occurs since the optimal temperature on the calibration data was learnt to be $T = 1$, which corresponds to not changing the base model at all. Even in such cases when temperature scaling changes the base model by a little, or not at all, top-label HB is shows improved performance.

Next, we discuss results with respect to conditional calibration, or the maximum calibration error (MCE). Guo et al. (2017) defined MCE with respect to confidence calibration, as follows:

$$\text{conf-MCE}(c,h) := \sup_{r \in \text{Range}(h)} \left| P(Y = c(X) \mid h(X) = r) - r \right|. \tag{13}$$

Conf-MCE suffers from the same issue illustrated in Figure 1. In Figure 1c, we looked at the reliability diagram within two bins. These indicate two of the values over which the supremum is taken in equation (13): these are the Y-axis distances (both less than 0.02) between the ★ markers and the $X = Y$ line for bins 6 and 10. On the other hand, the effective




(a) Optimized using focal loss.

(b) Optimized using Brier score.

*Figure 10.* Conf-ECE and TL-ECE of histogram binning and temperature scaling with deep nets on CIFAR-100. When assessed with respect to TL-ECE (solid lines in all plots), temperature scaling typically performs better than histogram binning. In fact, the performance of histogram binning is often worse than even the base model.

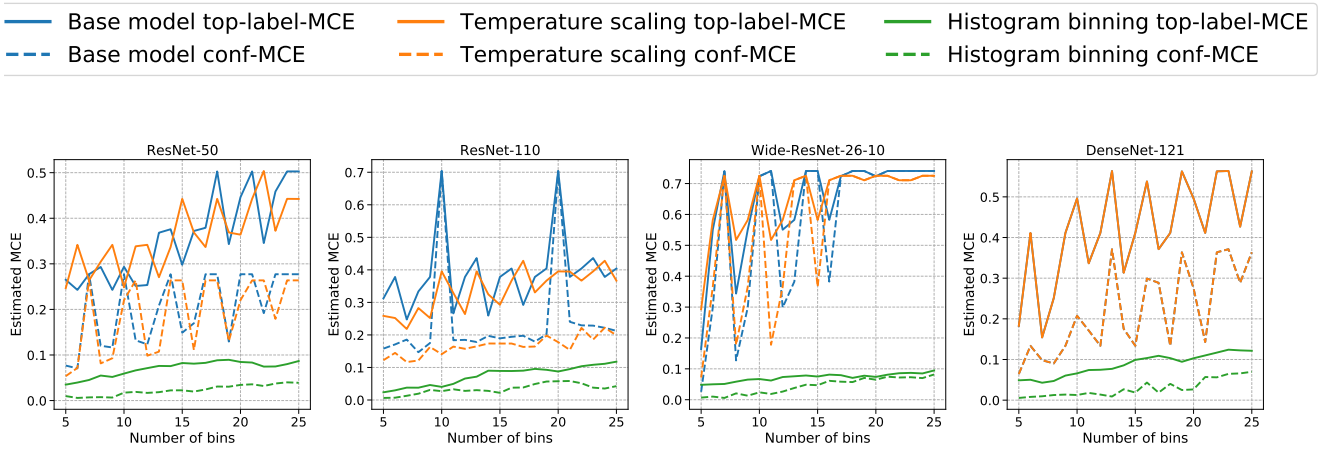
'maximum' miscalibration for bin 6 is roughly $0.15$ (for class 1), and roughly $0.045$ (for class 4), and the maximum should be taken with respect to these values across all bins. To remedy the underestimation of the effective MCE, we can consider the top-label-MCE, defined as follows:

$$\text{TL-MCE}(c,h) := \max_{l\in[L]} \sup_{r\in\text{Range}(h)} \left|P(Y=l \mid c(X)=l, h(X)=r) - r\right|. \tag{14}$$

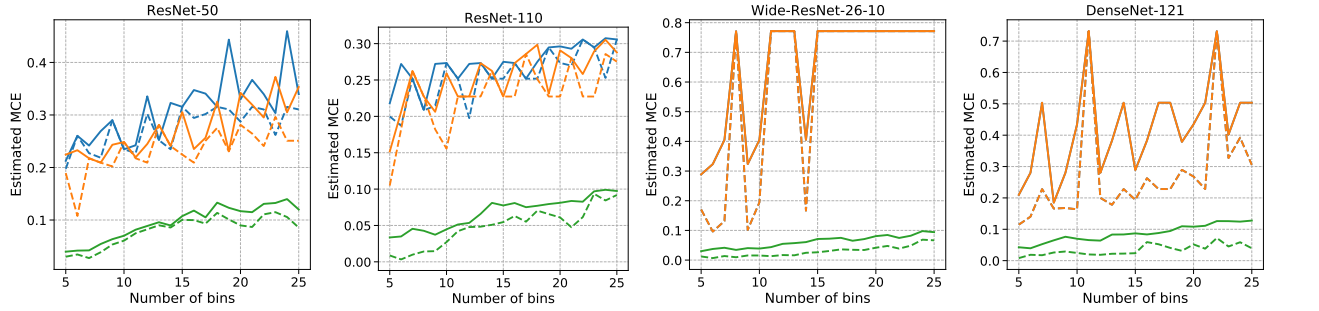
Interpreted in words, the TL-MCE assesses the maximum deviation between the predicted and true probabilities across all predictions and all classes. Following the same argument as in the proof of Proposition 1, it can be shown that for any $c, h$, conf-MCE$(c,h) \le$ TL-MCE$(c,h)$. The TL-MCE is closely related to conditional top-label calibration (Definition 1b). Clearly, an algorithm is $(\varepsilon,\alpha)$-conditionally top-label calibrated if and only if for every distribution $P$, $P(\text{TL-MCE}(c,h) \le \varepsilon) \ge 1-\alpha$. Thus the conditional top-label calibration guarantee of Theorem 1 implies a high probability bound on the TL-MCE as well. Since top-label HB ensures that each bin has roughly $k$ points, we expect that reasonable levels of calibration will be achieved across all bins. Consequently, we expect the MCE of top-label HB to be relatively small. This is confirmed empirically in the upcoming results.

Figure 11 shows conf-MCE and TL-MCE estimates of deep nets on CIFAR-10 for top-label HB, temperature scaling, and the base model. In Section 2, we had observed that top-label HB shows a slight but consistent improvement over temperature scaling for TL-ECE. For TL-MCE, top-label HB shows a *massive and consistent* improvement over temperature scaling across all architectures. Notice that for $B = 15$, the TL-MCE of temperature scaling is always larger than $0.2$. On the other hand, across different deep net architectures, loss functions, and numbers of bins, the TL-MCE of top-label HB is always below $0.1$. This is a powerful empirical result: for *every* test point $x \in \mathcal{X}$, the prediction $h(x)$ provided by top-label HB is at most $0.1$ away from the true probability $\mathbb{E}\left[Y = c(x) \mid c(X) = c(x), h(X) = h(x)\right]$.

Figure 12 shows conf-MCE and TL-MCE estimates of deep nets on CIFAR-100 for top-label HB, temperature scaling, and




(a) Optimized using focal loss.




(b) Optimized using Brier score.

*Figure 11.* Conf-MCE and TL-MCE of top-label histogram binning (HB) and temperature scaling (TS) with deep nets on CIFAR-10. HB shows multi-fold improvements over TS across architectures, loss functions, and numbers of bins. Unlike TS, the MCE is stable across different numbers of bins. For the Wide-ResNet-26-10 model optimized with Brier score, the base model has an estimated TL-MCE as well as conf-MCE of around 0.77. While TS did not recalibrate this model at all ($T = 1$), top-label HB achives a TL-MCE of around 0.065, more than a 10-fold improvement.

the base model. In Section 2, we had observed that temperature scaling performs better than top-label HB when assessed with respect to TL-ECE. For TL-MCE, we observe that the performance of top-label HB and temperature scaling is quite comparable. For $B = 15$, the TL-MCE of top-label HB is always below $0.3$. Further, the MCE of temperature scaling is quite unstable as the number of bins changes, the most instabilities occuring for the ResNet-50 and ResNet-100 plots with focal loss.

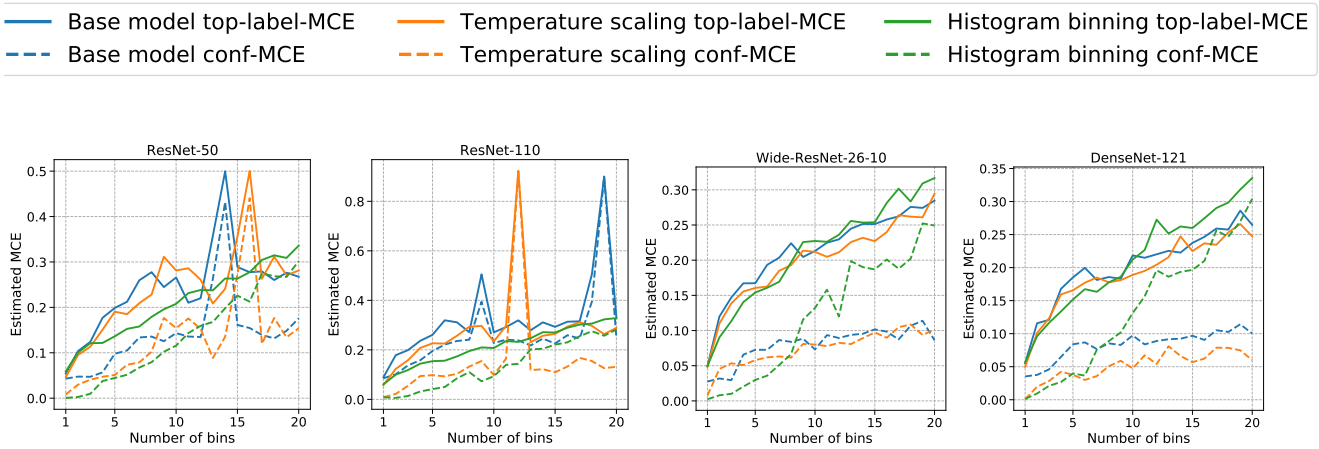
## E. Proofs

Proofs appear in separate subsections, in the same order as the corresponding results appear in the paper.

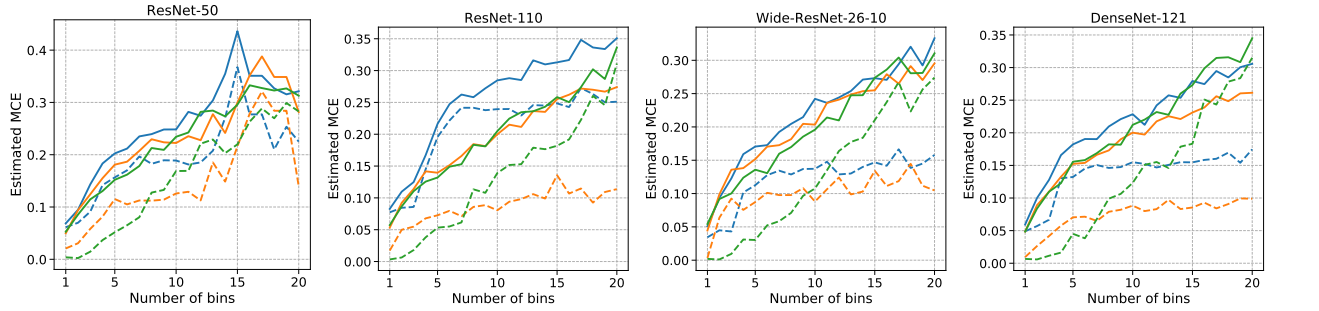
### E.1. Proof of Proposition 1

To avoid confusion between the the conditioning operator and the absolute value operator $|\cdot|$, we use $\mathrm{abs}\,(\cdot)$ to denote absolute values below. Note that,

$$\begin{aligned}
\mathrm{abs}\left(P(Y = c(X) \mid h(X)) - h(X)\right) &= \mathrm{abs}\left(\mathbb{E}\left[\mathbb{1}\left\{Y = c(X)\right\} \mid h(X)\right] - h(X)\right) \\
&= \mathrm{abs}\left(\mathbb{E}\left[\mathbb{1}\left\{Y = c(X)\right\} - h(X) \mid h(X)\right]\right) \\
&= \mathrm{abs}\left(\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Y = c(X)\right\} - h(X) \mid h(X), c(X)\right] \mid h(X)\right]\right) \\
&\leq \mathbb{E}\left[\mathrm{abs}\left(\mathbb{E}\left[\mathbb{1}\left\{Y = c(X)\right\} - h(X) \mid h(X), c(X)\right]\right) \mid h(X)\right] \\
&\qquad \text{(by Jensen's inequality)}
\end{aligned}$$




(a) Optimized using focal loss.




(b) Optimized using Brier score.

*Figure 12.* Conf-MCE and TL-MCE of top-label histogram binning (HB) and temperature scaling (TS) with deep nets on CIFAR-100. The TL-MCE of HB and TS is comparable, but HB is more stable across different numbers of bins.

$$= \mathbb{E}\left[\text{abs}\left(P(Y = c(X) \mid h(X), c(X)) - h(X)\right) \mid h(X)\right].$$

Thus,

$$\begin{aligned}
\text{conf-ECE}(c, h) &= \mathbb{E}\left[\text{abs}\left(P(Y = c(X) \mid h(X)) - h(X)\right)\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\text{abs}\left(P(Y = c(X) \mid h(X), c(X)) - h(X)\right) \mid h(X)\right]\right] \\
&= \mathbb{E}\left[\text{abs}\left(P(Y = c(X) \mid h(X), c(X)) - h(X)\right)\right] \\
&= \text{TL-ECE}(c, h).
\end{aligned}$$

□

### E.2. Proof of Theorem 1

The proof strategy is as follows. First, we use the bound of Gupta and Ramdas (2021, Theorem 4) (henceforth called the GR21 bound), derived in the binary calibration setting, to conclude marginal, conditional, and ECE guarantees for each $h_l$. Then, we show that the binary guarantees for the individual $h_l$'s leads to a top-label guarantee for the overall predictor $(c, h)$.

Consider any $l \in [L]$. Let $P_l$ denote the conditional distribution of $(X, \mathbb{1}\{Y = l\})$ given $c(X) = l$. Clearly, $D_l$ is a set of $n_l$ i.i.d. samples from $P_l$, and $h_l$ is learning a binary calibrator with respect to $P_l$ using binary histogram binning. The number of data-points is $n_l$ and the number of bins is $B_l = \lfloor n_l/k \rfloor$ bins. We now apply the GR21 bounds on $h_l$. First, we verify that the condition they require is satisfied:

$$n_l \geq k \lfloor n_l/k \rfloor \geq 2B_l.$$

Thus their marginal calibration bound for $h_l$ gives,

$$P\left(|P(Y = l \mid c(X) = l, h_l(X)) - h_l(X)| \leq \delta + \sqrt{\frac{\log(2/\alpha)}{2(\lfloor n_l/B_l \rfloor - 1)}} \mid c(X) = l\right) \geq 1 - \alpha.$$

Note that since $\lfloor n_l/B_l \rfloor = \lfloor n_l/\lfloor n_l/k \rfloor \rfloor \geq k$,

$$\varepsilon_1 = \delta + \sqrt{\frac{\log(2/\alpha)}{2(k-1)}} \geq \delta + \sqrt{\frac{\log(2/\alpha)}{2(\lfloor n_l/B_l \rfloor - 1)}}.$$

Thus we have

$$P\left(|P(Y = l \mid c(X) = l, h_l(X)) - h_l(X)| \leq \varepsilon_1 \mid c(X) = l\right) \geq 1 - \alpha.$$

This is satisfied for every $l$. Using the law of total probability gives us the top-label marginal calibration guarantee for $(c, h)$:

$$\begin{aligned}
&P(|P(Y = c(X) \mid c(X), h(X)) - h(X)| \leq \varepsilon_1) \\
&= \sum_{l=1}^{L} P(c(X) = l) P(|P(Y = c(X) \mid c(X), h(X)) - h(X)| \leq \varepsilon_1 \mid c(X) = l) \\
&\qquad \text{(law of total probability)} \\
&= \sum_{l=1}^{L} P(c(X) = l) P(|P(Y = l \mid c(X) = l, h_l(X)) - h_l(X)| \leq \varepsilon_1 \mid c(X) = l) \\
&\qquad \text{(by construction, if } c(x) = l, h(x) = h_l(x)) \\
&\geq \sum_{l=1}^{L} P(c(X) = l)(1 - \alpha) \\
&= 1 - \alpha.
\end{aligned}$$

Similarly, the in-expectation ECE bound of GR21, for $p = 1$, gives for every $l$,

$$\begin{aligned}
\mathbb{E}\left[|P(Y = l \mid c(X) = l, h_l(X)) - h_l(X) \mid c(X) = l|\right] &\leq \sqrt{B_l/2n_l} + \delta \\
&= \sqrt{\lfloor n_l/k \rfloor / 2n_l} + \delta \\
&\leq \sqrt{1/2k} + \delta.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[|P(Y = c(X) \mid c(X), h_l(X)) - h(X)|\right] \\
&= \sum_{l=1}^{L} P(c(X) = l) \mathbb{E}\left[|P(Y = l \mid c(X) = l, h_l(X)) - h_l(X)| \mid c(X) = l\right] \\
&\leq \sum_{l=1}^{L} P(c(X) = l)(\sqrt{1/2k} + \delta) \\
&= \sqrt{1/2k} + \delta.
\end{aligned}$$

Next, we show the top-label conditional calibration bound. Let $B = \sum_{l=1}^{L} B_l$ and $\alpha_l = \alpha B_l / B$. Note that $B \leq \sum_{l=1}^{L} n_l/k = n/k$. The binary conditional calibration bound of GR21 gives

$$P\left(\forall r \in \text{Range}(h_l), |P(Y = l \mid c(X) = l, h_l(X) = r) - r| \leq \delta + \sqrt{\frac{\log(2B_l/\alpha_l)}{2(\lfloor n_l/B_l \rfloor - 1)}} \mid c(X) = l\right)$$

$$\geq 1 - \alpha_l.$$

Note that

$$\begin{aligned}
\sqrt{\frac{\log(2B_l/\alpha_l)}{2(\lfloor n_l/B_l \rfloor - 1)}} &= \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n_l/B_l \rfloor - 1)}} \\
&\leq \sqrt{\frac{\log(2n/k\alpha)}{2(\lfloor n_l/B_l \rfloor - 1)}} && \text{(since } B \leq n/k\text{)} \\
&\leq \sqrt{\frac{\log(2n/k\alpha)}{2(k-1)}} && \text{(since } k \leq \lfloor n_l/B_l \rfloor\text{).}
\end{aligned}$$

Thus for every $l \in [L]$,

$$P(\forall r \in \text{Range}(h_l),\ |P(Y = l \mid c(X) = l, h_l(X) = r) - r| \leq \varepsilon_2) \geq 1 - \alpha_l.$$

By construction of $h$, conditioning on $c(X) = l$ and $h_l(X) = r$ is the same as conditioning on $c(X) = l$ and $h(X) = r$. Taking a union bound over all $L$ gives

$$\begin{aligned}
&P(\forall l \in [L], r \in \text{Range}(h), |P(Y = l \mid c(X) = l, h(X) = r) - r|) \leq \varepsilon_2) \\
&\qquad \geq 1 - \sum_{l=1}^{L} \alpha_l = 1 - \alpha.
\end{aligned}$$

Finally, note that if for every $l \in [L], r \in \text{Range}(h)$,

$$|P(Y = l \mid c(X) = l, h(X) = r) - r|) \leq \varepsilon_2,$$

then also

$$\text{TL-ECE}(c, h) = \mathbb{E}\left[|P(Y = c(X) \mid h(X), c(X)) - h(X)|\right] \leq \varepsilon_2.$$

This proves the high-probability bound for the TL-ECE. □

*Remark* 1. Gupta and Ramdas (2021) proved a more general result for general $\ell_p$-ECE bounds. Similar results can also be derived for the suitably defined $\ell_p$-TL-ECE. In particular, it can be shown that with probability $1-\alpha$, the TL-MCE of $(c, h)$ which can be seen as the $\ell_\infty$-TL-ECE of $(c, h)$ is bounded by $\varepsilon_2$. (TL-MCE is defined in Appendix D.2, equation (14).)

### E.3. Proof of Theorem 2

We use the bound of Gupta and Ramdas (2021, Theorem 4) (henceforth called the GR21 bound), derived in the binary calibration setting, to conclude marginal, conditional, and ECE guarantees for each $h_l$. This leads to the class-wise results as well.

Consider any $l \in [L]$. Let $P_l$ denote the distribution of $(X, \mathbb{1}\{Y = l\})$. Clearly, $D_l$ is a set of $n$ i.i.d. samples from $P_l$, and $h_l$ is learning a binary calibrator with respect to $P_l$ using binary histogram binning. The number of data-points is $n$ and the number of bins is $B_l = \lfloor n/k_l \rfloor$ bins. We now apply the GR21 bounds on $h_l$. First, we verify that the condition they require is satisfied:

$$n \geq k_l \lfloor n/k_l \rfloor \geq 2B_l.$$

Thus the GR21 marginal calibration bound gives that for every $l \in [L]$, $h_l$ satisfies

$$P\left(|P(Y = l \mid h_l(X)) - h_l(X)| \leq \delta + \sqrt{\frac{\log(2/\alpha_l)}{2(\lfloor n/B_l \rfloor - 1)}}\right) \geq 1 - \alpha_l.$$

The class-wise marginal calibration bound of Theorem 2 follows since $\lfloor n/B_l \rfloor = \lfloor n/\lfloor n/k_l \rfloor \rfloor \geq k_l$, and so

$$\varepsilon_l^{(1)} \geq \delta + \sqrt{\frac{\log(2/\alpha_l)}{2(\lfloor n/B_l \rfloor - 1)}}.$$

Next, the GR21 conditional calibration bound gives for every $l \in [L]$, $h_l$ satisfies

$$P\left(\forall r \in \text{Range}(h_l), |P(Y = l \mid h_l(X) = r) - r| \leq \delta + \sqrt{\frac{\log(2B_l/\alpha_l)}{2(\lfloor n/B_l \rfloor - 1)}}\right) \geq 1 - \alpha_l.$$

The class-wise marginal calibration bound of Theorem 2 follows since $B_l = \lfloor n/k_l \rfloor \leq n/k_l$ and $\lfloor n/B_l \rfloor = \lfloor n/\lfloor n/k_l \rfloor \rfloor \geq k_l$, and so

$$\varepsilon_l^{(2)} \geq \delta + \sqrt{\frac{\log(2B_l/\alpha_l)}{2(\lfloor n/B_l \rfloor - 1)}}.$$

Let $k = \max_{l \in [L]} k_l$. The in-expectation ECE bound of GR21, for $p = 1$, gives for every $l$,

$$\begin{aligned}
\mathbb{E}\left[\text{binary-ECE-for-class-}l\ (h_l)\right] &\leq \sqrt{B_l/2n_l} + \delta \\
&= \sqrt{\lfloor n/k_l \rfloor /2n} + \delta \\
&\leq \sqrt{1/2k_l} + \delta \\
&\leq \sqrt{1/2k} + \delta.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\text{CW-ECE}(c,h)\right] &= \mathbb{E}\left[L^{-1}\sum_{l=1}^{L} \text{binary-ECE-for-class-}l\ (h_l)\right] \\
&\leq L^{-1}\sum_{l=1}^{L}(\sqrt{1/2k} + \delta) \\
&= \sqrt{1/2k} + \delta,
\end{aligned}$$

as required for the in-expectation CW-ECE bound of Theorem 2. Finally, for the high probability CW-ECE bound, let $\varepsilon = \max_{l \in [L]} \varepsilon_l^{(2)}$ and $\alpha = \sum_{l=1}^{L} \alpha_l$. By taking a union bound over the the conditional calibration bounds for each $h_l$, we have: with probability $1 - \alpha$, for every $l \in [L], r \in \text{Range}(h)$,

$$|P(Y = l \mid c(X) = l, h(X) = r) - r|) \leq \varepsilon_l^{(2)} \leq \varepsilon.$$

Thus, with probability $1 - \alpha$,

$$\text{CW-ECE}(c,h) = L^{-1}\sum_{l=1}^{L} \mathbb{E}\left[|P(Y = l \mid h_l(X)) - h_l(X)|\right] \leq \varepsilon.$$

This proves the high-probability bound for the CW-ECE. □